Crowdsourcing the OCR Ground Truth of a German and French Cultural Heritage Corpus

Clematide, Simon ; Furrer, Lenz ; Volk, Martin

Abstract: Crowdsourcing approaches for post-correction of OCR output (Optical Character Recognition) have been successfully applied to several historical text collections. We report on our crowd-correction platform Kokos, which we built to improve the OCR quality of the digitized yearbooks of the Swiss Alpine Club (SAC) from the 19th century. This multilingual heritage corpus consists of Alpine texts mainly written in German and French, all typeset in Antiqua font. Finding and engaging volunteers for correcting large amounts of pages into high quality text requires a carefully designed user interface, an easy-to-use workflow, and continuous efforts for keeping the participants motivated. More than 180,000 characters on about 21,000 pages were corrected by volunteers in about 7 months, achieving an OCR ground truth with a systematically evaluated accuracy of 99.7 on the word level. The crowdsourced OCR ground truth and the corresponding original OCR recognition results from Abbyy FineReader for each page are available as a resource for machine learning and evaluation. Additionally, the scanned images (300 dpi) of all pages are included to enable tests with other OCR software.

Originally published at:
Clematide, Simon; Furrer, Lenz; Volk, Martin (2018). Crowdsourcing the OCR Ground Truth of a German and French Cultural Heritage Corpus. Journal for Language Technology and Computational Linguistics (JLCL), 33(1):25-47.

Simon Clematide, Lenz Furrer, Martin Volk

Crowdsourcing the OCR Ground Truth of a German and French Cultural Heritage Corpus

Abstract

Crowdsourcing approaches for post-correction of OCR output (Optical Character Recognition) have been successfully applied to several historical text collections. We report on our crowd-correction platform Kokos, which we built to improve the OCR quality of the digitized yearbooks of the Swiss Alpine Club (SAC) from the 19th century. This multilingual heritage corpus consists of Alpine texts mainly written in German and French, all typeset in Antiqua font. Finding and engaging volunteers for correcting large amounts of pages into high quality text requires a carefully designed user interface, an easy-to-use workflow, and continuous efforts for keeping the participants motivated. More than 180,000 characters on about 21,000 pages were corrected by volunteers in about 7 months, achieving an OCR ground truth with a systematically evaluated accuracy of 99.7 % on the word level. The crowdsourced OCR ground truth and the corresponding original OCR recognition results from Abbyy FineReader for each page are available as a resource for machine learning and evaluation. Additionally, the scanned images (300 dpi) of all pages are included to enable tests with other OCR software.

1 Introduction

Crowdsourcing approaches for post-correction of Optical Character Recognition (OCR) output have been successfully applied to several historical text collections (Holley, 2009b; DTAQ, 2016). We report on our crowd-correction platform Kokos,[1] which we built to improve the text quality of the digitized yearbooks of the Swiss Alpine Club (SAC)[2] from the 19th century. This multilingual heritage corpus consists of Alpine texts mainly written in German and French, all typeset in Antiqua font.

Finding and engaging volunteers for correcting large amounts of automatically OCRed pages into high quality text requires a carefully designed user interface, an easy-to-use workflow, and continuous efforts for keeping the participants motivated.

The scanned images, the uncorrected output of a standard OCR software and the high-quality text corrected by our crowd are a valuable resource.[3] It can be used for

[1] http://kokos.cl.uzh.ch
[2] http://www.sac-cas.ch
[3] http://pub.cl.uzh.ch/purl/OCR19thSAC

extracting heritage lexicons covering 19th century German in particular, or for training as well as testing automatic OCR error correction systems.

In the following section, we introduce our multilingual corpus and describe the process of its digitization. We report on our efforts in building and maintaining a crowd-correction platform and compare them to other work in the field. In Section 3, we analyze and evaluate the corrections performed by the volunteer collaborators. The released resource is described in the last subsection.

2 Materials and Methods

2.1 Corpus Data

In the Text+Berg project[4] we digitized the yearbooks of the Swiss Alpine Club (SAC) from 1864 until today (henceforth SAC corpus) for building a multilingual heritage corpus of Alpine texts (Göhring & Volk, 2011).

In this paper we focus on the yearbooks from the 19th century. Including tables of content and index pages, the books from 1864 to 1899 amount to 21,246 pages with around 304,000 sentences and 6.3 million tokens (before correction). This is about 16 % of our complete SAC corpus.

Thematically, the corpus contains detailed mountaineering and travel reports (mostly from Switzerland, but also from abroad), historical and biological articles (flora and fauna of the Alps), geological and geographical studies (including frequent glacier observations), linguistic articles (e. g. on language boundaries in the Alps), and protocols of the annual club meetings. The text contains a huge number of proper names, geographical names, and Latin botanical names.

Our statistical sentence-based language identification (Dunning, 1994)[5] assigns 5.5 million tokens to German and 0.74 million tokens to French. See Figure 3 for the distribution of these languages across yearbooks. Additionally, there are a few thousand tokens in English (mostly book and article titles), Italian, Swiss German, and Romansh,[6] but note that these numbers do not reflect code-switching within sentences (Volk & Clematide, 2014).

2.2 OCR

All the yearbooks from 1864 until 2000 have been collected in printed form. From 2001 until 2009 the SAC has provided us with PDF files, and since 2011 the SAC generates structured XML files directly out of their authoring system.

We obtained the first 10 yearbooks as leather-bound copies. Through collaboration with the Austrian Academy Corpus (AAC) group in Vienna, we scanned them without destroying them. All yearbooks from 1874 until 2000 were cut open so that we were

[4] `http://textberg.ch`

[5] We use M. Piotrowski's PERL reimplementation Lingua::Ident.

[6] Most of the 384 sentences (the vast majority) of the 19th century that were automatically classified as Romansh were in fact Latin, French, toponyms, or OCR errors.

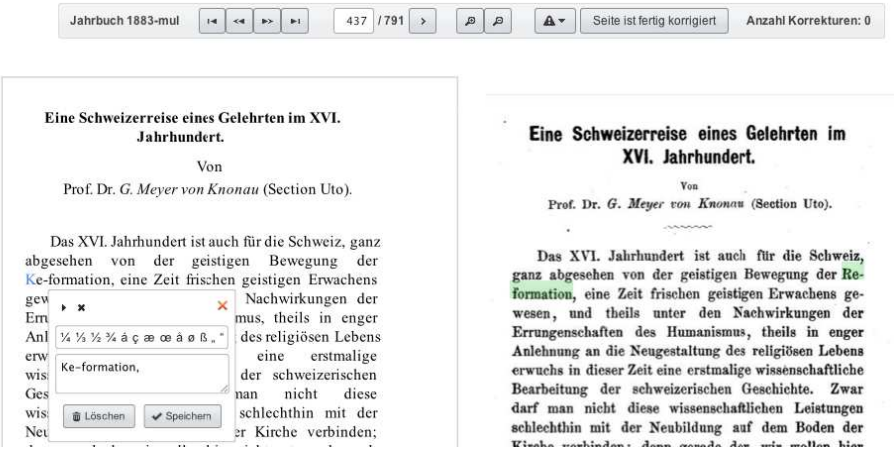

Figure 1: A book page in Kokos: synoptic view of the editable text on the left and the facsimile image on the right. Note the small edit window within the text and the corresponding highlighted word in the facsimile.

able to use a scanner with paper feed. From 1957 onwards, the SAC has published parallel French and German versions of the yearbooks, both of which we processed in the same manner.

After scanning all book pages with 300 dpi, we used the OCR software Abbyy FineReader Pro 7 to convert the images to text (selecting the recognition languages German, French and Italian). This led to mixed text recognition results. The text on some pages was recognized excellently whereas other pages contained a multitude of OCR errors.

Our initial idea was to manually correct these errors in the OCR system since it preserves the mapping between words recognized in the text and the corresponding position on the page. But we soon realized that manual correction is very time-consuming even when working on well-recognized yearbooks of the 20th century. It is prohibitively time-consuming for the yearbooks of the 19th century, where recognition accuracy is inferior because of (a) words that are unknown to the OCR system lexicon (foreign words, old German spellings, toponyms, special terms used in mountaineering, person names, dialect words), (b) special characters (fraction glyphs, Greek letters, old symbols), (c) stains on the paper and curved pages. Generally, the corpus contains many challenges for OCR and Optical Layout Recognition, such as tables, mathematical formulae, spaced type, or words in images.

We investigated various means of improving the OCR quality and correcting OCR errors automatically (Volk, Marek, & Sennrich, 2010). There are only few ways in which a commercial OCR system can be tuned. The most obvious way is to add

"unknown" words to its lexicon. In order to extend the coverage of the built-in lexicon, we collected words with old German spelling patterns (e. g. *acceptiren, acceptieren, Mittheilung*) and also added the names of 4000 Swiss mountains and cities. This led to some improvements of the OCR quality but a multitude of seemingly random OCR errors persisted.[7]

Then we experimented with two ways of automatic error correction. First we employed a second OCR system (OmniPage) and compared the output of the two systems (Volk et al., 2010). Wherever they disagreed we checked with a German morphological system (Gertwol, see Koskeniemmi & Haapalainen, 1996) whether both words were known German words. If so, then we chose the word that occurred more frequently in our corpus. If only one of the words was known, then this was the obvious choice. If none of the words was known, then we trusted Abbyy FineReader as the more reliable system. This method also led to a small reduction of errors.

Finally we experimented with automatic error correction based on character similarities of words. If an unknown word deviates only in one or two characters from another known word which frequently occurs in our corpus, then we automatically substitute the unknown word with the known word. This method is similar to grammar checking as used in popular text processing software, but needs to work with high precision since human intervention (i. e. manual choice of the correct option) is not possible given the large amounts of text. Therefore we applied this method only for words with a length of more than 15 characters. After all these efforts many spurious OCR errors persisted.

2.3 Crowd Correction

It became obvious that we can only achieve a clean corpus if we organize a large distributed effort for correcting OCR errors via a crowd of volunteers. Therefore, we built the collaborative web-based correction system *Kokos*.[8] Kokos is based on the wiki idea and is technically built on top of *PmWiki*.[9] The initial OCR content in Kokos consists of the original Abby FineReader output of all yearbooks of the 19th century, as one of our goals was the assessement of quality of crowdcorrection.

2.3.1 User Interface

We modified the wiki such that it displays the OCRed text of a page and the scan image side by side (see Figure 1). The text is an HTML export from the OCR software, and the layout, paragraphs and font sizes resemble the facsimile.

In the recognized text, each word is a clickable and editable unit. While reading through the text, Kokos correctors can simply click on faulty words in order to open a small editing window (Figure 1). In this window they can modify the word and save the correction. Quick access buttons help to insert frequent incorrectly recognized special

[7] Holley (2009a) comes to similar conclusions.
[8] `http://kokos.cl.uzh.ch`
[9] `http://www.pmwiki.org`

characters, e. g. æ, ß, ¼, or Greek letters. The corrected word immediately becomes visible in the text.

In addition to the correction of characters within a word, three generic operations on the level of one or more adjacent words are frequently needed. First, a delete button removes spurious tokens typically caused by dirt or stains on the page. A second button joins incorrectly split tokens into the edit window, for instance, in the case of spaced type, which was often used to highlight certain words in the 19th century. Third, inadvertently connected words can be split by inserting a blank character.

When the editing window is open or when the user hovers over a word, the corresponding rectangle in the facsimile is highlighted. This is an important and motivating feature that allows the user to quickly spot and doublecheck a suspicious word in the image. The positions of each word were computed by the OCR system during recognition. These coordinates provide the alignment between each word in the text and the corresponding area in the image.

In order to draw the reader's attention to words where the OCR software had low recognition confidence (that is, potential OCR errors), a blue font color was used. Unfortunately, the confidence values of the software were not as reliable as we had hoped, and therefore not as helpful for guiding the human correctors.

In addition to correcting OCR errors, we asked the users to perform *dehyphenation*, i. e. recomposing words that were hyphenated at a line break.

2.3.2 Workflow

In order to attract correctors to work on the task it is important to make initial access as easy as possible. In Kokos we allowed all interested persons to read through the text by browsing and searching. It was then an easy step to register with user name, password and email address in order to sign up as a volunteer corrector. The downside of this is that we know very little about our correctors.

In order to achieve consistent corrections, we provided a set of concise guidelines with typical examples. Additionally, we curated a list of frequently asked questions (FAQ), which was updated according to the problems which our correctors reported. It probably would have been a good idea to introduce the task and the correction guidelines with a short tutorial video.

Users can access the text through a table of contents sorted by yearbook, or a text search, or a "Quick Start" button that leads to a page without or only a few corrections, or an overview of finished and unfinished pages. Especially in the final phase of correction, this view guided our volunteers to correct yearbooks completely. Our basic workflow is "correct errors while reading a text of interest". That crowd correction is driven by curiosity became obvious to us when we noticed that all reports about accidents in the mountains were corrected early on. They are exciting and thrilling. On the other hand, articles about the geology of the Alps with many technical terms (like e.g. *Quarzporphyr, Kreideprotogine, paläozoische Granite* in 1889) were left until the end.

By clicking on a button, users can mark a page as finished when they consider the text carefully corrected. This button will also automatically advance the view to the next page. Other users can still apply corrections to "finished" pages if need be. We had pondered over whether to lock a page after a user has marked it as finished. The advantage would have been that the page then cannot be affected any more by vandalism. We decided against this automatic locking in order to allow for post-corrections and to send a signal of trust to our contributors. This worked fine.

Kokos also supports an orthogonal workflow via global search and replace, which includes a keyword-in-context view of the search results with facsimile image snippets of the search word (see Figure 2). This speeds up the correction of repeated recognition errors. In order to prevent users from introducing damage by accidental mass replacements, we limited the amount of global replacements to 15 hits per user interaction.

On each page, the correctors were reminded to preserve the spelling of the original text,[10] even if it deviated from modern orthography, or even if they encountered one of the very rare printing errors in our carefully typeset books.

2.3.3 Crowd Management

In January 2014, the SAC monthly magazine LES ALPES, DIE ALPEN, LE ALPI (in all three language versions: French, German and Italian) published a call for volunteer helpers to correct our SAC heritage yearbooks. Dozens of users registered in Kokos and started to contribute within days. After 7 months our active crowd had finished correcting all of the 21,000 pages. We observed a performance pattern which seems to be typical for crowd correction (Holley, 2009b, 15): there were not thousands of volunteers doing tiny bits of work (typical for paid micro-work crowdsourcing), but there was a small crowd of dedicated correctors doing most of the work.

Our correctors were cooperative and reliable, for instance, regarding marking pages as corrected, and we never had to deal with vandalism. Our initial fears that we needed to invest a lot of time to monitor the correction quality, or that a double correction of all pages would be necessary in order to achieve the envisaged quality turned out to be unsubstantiated. This is very much in line with Holley's tip 14 "Assume volunteers will do it right rather than wrong" (Holley, 2010).

In order to keep the top performers motivated and to give them feedback, a user ranking based on the number of corrections proved to be useful. In our opinion, this kind of gamification is sufficient for volunteers who are inherently interested in a task. For community building, we regularly sent emails to the correctors once a month, informing them about progress and system improvements.

Via social media buttons which we had integrated into each Kokos web page, the users could promote interesting pages to common social media channels. However, this

[10] The most important deviations from modern orthography are "c" instead of "z" or "k", "i" instead of "ie" (*acceptiren*, modern form: *akzeptieren*), "th" instead of "t" (*Thal*).

Figure 2: Search result in KWIC view with facsimile snippets for each hit for quick verification.

feature was not used a lot by our correctors, and therefore did not help to attract more volunteer workers.

Even though our crowd correction initiative was advertised in all Swiss language regions, the French texts in our collection were corrected late. We suspect that one of the reasons was that we only offered a German user interface which made the Kokos system foreign to French speakers. It is clearly important that the user interface including the guidelines and the FAQ must be provided in the languages of all targeted contributors.

2.4 Related Work

In order to achieve high quality in the retrodigitization of printed historical text material, there are two viable options (DFG, 2009): (a) manual transcription, or (b) OCR and post-correction.

In the case of manual transcription, independent double-keying by non-native speakers achieves the highest quality (typically, the contracted accuracy on character level is higher than 99.95%), but is most expensive.[11] For historical German, Haaf, Wiegand, and Geyken (2013) confirmed these transcription accuracy numbers in their systematic and representative evaluation on texts from 1780 to 1899 taken from the *Deutsches Textarchiv (DTA)*.[12] A sample of 7,208 pages with 9.9 million characters in total was proofread in DTA's quality assurance platform DTAQ (2016) and 830 transcription errors were revealed. This translates into an overall character-level accuracy of 99.99%. Surprisingly, the accuracy of Antiqua and Gothic typeface was roughly the same.

In the case of OCR, post-correction can be done automatically or manually, for instance, by crowd correction. In the remainder of this section, we discuss relevant manual approaches and initiatives related to our work.

[11] Offshore double-keying costs between 0.4 and 0.8 euros per 1,000 keystrokes, depending on structural markup and typeface (Piotrowski, 2012). An accuracy higher than 99% is standard (Long, 1993).
[12] `http://www.deutschestextarchiv.de`

2.4.1 Crowd Correction

The *Distributed Proofreaders* web site,[13] founded in 2000 by Charles Frank in order to assist the Project Gutenberg in the digitization of Public Domain books, is probably the first crowd-correction initiative, and still active with several thousands of volunteers. The users proofread the OCRed raw text in a simple textual input form while reading through the facsimile. No visual synchronization between the transcribed words and their image location is available. Proofreading is done page per page in two separate rounds by two different proofreaders, optionally, a third round can be applied. In contrast to independent double-keying, these rounds follow each other. The site provides an interesting spell-checking correction mode that presents a view of the text where only words unknown to the spellchecker are editable and, in this way, guides the proofreading process. Book-specific white lists of known and verified word forms can be updated in this mode during the correction in order to adapt the spellchecker to the vocabulary of a text. A qualification system based on the amount of accomplished corrections and successfully passed quizzes concerning the guidelines regulates the type of work a volunteer is allowed to perform. Different statistics monitor the progress of the projects and the individual contributors. The user interface of the website is complex, which partly is due to the fact that the site also includes functionality for formatting the proofread e-books.

Wikisource,[14] another long-term volunteer crowd correction infrastructure, was founded in 2003 and has hundreds of active members. Its German and French instances contain several hundred thousand public domain German and French pages (books and single-leaf prints), many from the 19th century. The technical backbone of every wikisource site is a mediawiki plugin that displays scans and OCRed text side-by-side. No visual synchronization between the transcribed words and their image location is available. Wikisource aims at producing corrected material that satisfies scientific citability and needs. Wiki markup can be used directly in the proofreading phase in order to render some of the typographic layout, for instance spaced type. A page must be proofread in sequence by two different correctors in order to be considered as validated. The correction workflow is openly managed by wiki tags that are set by the users, however, validated pages are protected against further edits, and further corrections must be requested by the wiki discussion pages. On the *Wikisource* as well as on the *Distributed Proofreaders* web site, anyone can import new scanned text material for correction.

The *reCAPTCHA* system (von Ahn, Maurer, McMillen, Abraham, & Blum, 2008) has earned early fame for hiding crowdsourcing effort in OCR correction behind an access system to websites. Users are shown two artificially distorted image snippets where one is known to the system and used for preventing automated abusive access to a website. The other is unknown[15] and its text content will be determined by a

[13] `http://www.pgdp.net`

[14] `https://wikisource.org`

[15] Actually, the selection criterion for these words is that two different OCR systems suggested two different words.

majority vote of many contributors. Of course, users do not know which word is known and which is unknown. An evaluation on a sample of 50 articles (24,080 words) of the New York Times archive from 1860 to 1970 revealed an accuracy of 99.1% on the word level. This is a large improvement over the standard OCR software word accuracy of 83.5%, and very close to the 99.2% accuracy of the double-keyed transcription in its initial state. The final ground truth was produced by carefully comparing every difference between the manual transcription and the reCAPTCHA output.

The National Library of Australia has set up *trove*,[16] a system for crowd correction of OCRed historical newspapers (Holley, 2009a). The main goal of this initiative is to have the corrected text content accessible for fulltext search, therefore, typographical formatting information is not preserved. The guidelines also explicitly allow for corrections of obvious typesetting errors in the facsimile, however, they encourage the users to add corresponding comments. The user interface has to deal with the complex newspaper column layout. Corrections are applied line by line to segmented articles and the user interface dynamically highlights the corresponding line in the facsimile. There is no proofreader workflow defined on the level of articles or pages, correctors can change any text anytime.

An important technical measure of trove for avoiding vandalism is the transparency of user edits: recent edit operations are streamed "live" in the user interface and any user can inspect the recent corrections of any other user. If users detect large amounts of spam or malicious corrections, they can request a roll back. Small fonts in combination with low paper and print quality of historic newspapers often produce bad OCR output, even with the best software available. From 2008 to the end of 2016, almost 220 million lines have been corrected manually. At the end of 2016, there were about 46,000 registered users, but many of them contributed only few corrections, but few correctors contributed a lot. The hall-of-fame reveals that the top ten volunteers have corrected 4.2, 2.9, 2.7, ..., 1.4 million lines by the end of 2016, which amounts to 21.4 millions in total and almost 10% of all corrections.

Commercial platforms for digitization and document collection management such as *Veridian*[17] have also successfully integrated user correction of digitized historical newspapers. Veridian is in use by several large libraries around the world. For instance, the *California Digital Newspaper Collection*[18] counted 7.45 million lines corrected by 2,582 users at the end of 2016. Again, the distribution of corrections per user is extreme: the top ten volunteers produced 4.06 millions (54.5%) of all corrections. The user interface and workflow is similar to *trove*; some functionality is missing though, such as the addition of lines that were not recognized at all by the OCR engine. Rose Holley's blog entry (Holley, 2013) lists five US historical newspapers that employ crowdsourcing for OCR corrections, as well as an Australian, Finnish, Russian, and Vietnamese one. The amount of newspaper pages in combination with the quality of the OCR output

[16] http://trove.nla.gov.au
[17] http://www.veridiansoftware.com
[18] http://cdnc.ucr.edu/cgi-bin/cdnc

make volunteer crowd correction a cost-effective instrument for improving access to these heritage data.

The *Deutsches Textarchiv (DTA)* (Geyken & Gloning, 2015), a large philological archive of 15th–19th-century German texts, includes a web-based quality assurance platform which is open to anyone (Haaf et al., 2013). Although many texts have been transcribed manually by double-keying and are almost free of errors, recently more OCRed texts have been added to the archive.[19] The document representation in the DTA is a highly structured XML format (Haaf, Wiegand, & Geyken, 2014), which requires specific expertise. Therefore, the correction workflow for volunteers is more restricted and adopts the paradigm of issue tickets known from software development. If a user detects a transcription error, he opens a ticket linked to the faulty words, describes the type of error in a form (which covers other issues as well, for instance, formatting or structural problems), and inserts the corrected words in a text field of the form. Each ticket is then resolved by a DTA staff member. This procedure for correcting text errors is more cumbersome for the user, slower than the aforementioned workflows, and probably not suited for the initial correction of raw OCR output. However, it guarantees the preservation of the high philological quality standards of the project. The correction workflow is based on pages and the user has to explicitly mark a page as corrected; the platform distinguishes two types of transcription validation, (a) a confirmation that the extracted text has been read carefully and no text problems have been found, and (b) a confirmation that the extracted text corresponds to the shown facsimile. No visual synchronization between the transcribed words and their image location is available.

The *PoCoTo* system for postcorrection of OCRed historical text comprises several tools and a web-based interface with an interactive workflow whose efficiency has been attested by a small user study (Vobl, Gotscharek, Reffle, Ringlstetter, & Schulz, 2014).[20] An interesting feature is the interlinear-like view where facsimile snippets of individual tokens and their recognized text are presented in reading order. The system suggests correction candidates computed in a corpus-based unsupervised manner (Reffle, 2011). The interface also offers batch correction of precomputed error series (e. g. $u \rightarrow n$) in a concordance view.

Citizen science web sites such as *crowdcrafting*[21] offer a PDF transcription task template which can be used for hosting small-scale projects.

Our review shows that crowd correction for books typically works on the level of pages. For newspapers, corrections are typically applied on the level of individual lines in the context of an article. Yet another approach for involving the crowd into OCR correction is reported by Wang, Wang, and Chen (2013) for ancient Chinese books. They first extract graphically similar Chinese characters and present them to the users in a row for quick verification. This reduces the correction task to the question whether

[19] Some of them were taken from the German wikisource site in their corrected form.
[20] `https://github.com/cisocrgroup/PoCoTo`
[21] See `http://www.crowdcrafting.org`, which is based on the popular *pybossa* crowdsourcing framework.

all logograms in a row are the same. A prototype with a similar approach for old Venice manuscripts has been explored by Simeoni, Mazzei, and Kaplan (2014).

Chrons and Sundell (2011) present *Digitalkoot*,[22] a gamification-based system for correcting OCR errors in old Finnish newspapers typeset in Gothic font. The words are taken out of context and inserted into simple games. The authors monitored the activities for the initial two months, in which 4,800 persons played the games and completed 2.5 million microtasks. This was the result of heavy media coverage with more than 30 newspaper articles and some TV programs reporting on the project. The authors remark that a small percentage of users provided one third of the work, therefore showing a similar user behavior compared to volunteers. The quality of the crowd corrections was very high and improved the text from 85 % word accuracy to over 99 %. At the end of the project after 22 months, 8 million microtasks had been solved by the gamers, however, this did not result in 8 million corrected tokens. Several gamers had to agree on a transcription before it will be accepted. Additionally, many known items had to be presented in order to identify "cheaters". Therefore, Kettunen (2016) concludes that "this approach is clearly not feasible" for the correction of the 837 million words in this corpus and advocates improved automatic OCR post-correction.

Seidman, Flanagan, Rose-Sandler, and Lichtenberg (2016) describe another OCR correction initiative set up as purposeful gaming.[23] The basic idea is similar to reCAPTCHA: words which have been recognized differently by two independent OCR systems are presented to the crowd. The key challenge in the gamification of OCR corrections is the question of how to decide whether a user contribution for an "unknown" word should be rewarded as correct or not. Seidman et al solve the problem by decoupling the reward to the player from the decision on the ground truth. The player is rewarded if his/her contribution exactly matches one of the OCR suggestions, or if it exactly matches an accepted contribution created earlier by another player.[24] New contributions by players are only accepted in the system if they match the common substrings of the OCR systems. The ground truth for an unknown word is determined by a threshold of how many times an exact match was found for a contribution. Although the professionally designed game worked technically and even received an award in the field of purposeful gaming, Seidman et al. (2016) had to concede that this approach does not deliver the needed amount of corrections.

The main differences between our platform and the ones mentioned above are the following. First, our system is not practical for documents with complex layout such as newspapers. Second, from a technical point of view our platform is simple, however, it relies on specific HTML output formats produced by the OCR software. Third, instead of trying to create an artificial gaming setup where volunteers would only see single words in isolation, it is important for us to let them read historical documents while correcting OCR errors.

[22] `http://www.digitalkoot.fi`

[23] A working version of the game is online under `http://smorballgame.org`.

[24] The game logic tries to minimize false negatives at the cost of allowing the gamer to be rewarded for false positives. Seidman et al. (2016) estimate the false negative rate to be 3%.

In summary, it is safe to state that crowd correction by motivated volunteers is the best strategy to correct annoying OCR errors if your budget does not allow for paid work. Gamification cannot help to correct large amounts of OCR data, even when the input for the correction is reduced to the material where OCR systems disagree. Crowd correction of a large amount of text in a language with a large group of speakers will generally work better; however, even then typically only very few volunteers are doing most of the work. The example of Trove newspaper correction shows that volunteers are also helpful for correcting OCR output with a much lower quality than ours. As a positive side effect, involving a crowd of users is also a good way to disseminate the digitized cultural heritage material.

3 Results

We investigated our crowdsourced corrections in two ways: with a quantitative analysis of the modifications, which is detailed in Section 3.1, and by evaluating the quality of the corrected texts in a representative sample that was checked separately by two persons (Section 3.2).

3.1 Amount of Corrections

For assessing the amount of corrections, we compared two snapshots of the texts, taken at the start and at the end of the correction phase. We determined the amount of corrections by means of the modification rate between the two versions, which is the character edit distance (Levenshtein, 1966) divided by the length of the corrected text. We computed the modification rate with the ISRI frontiers toolkit (Rice, 1996), which is meant for analyzing modifications in OCR text. Across the entire corpus, the edit distance sums up to a total of almost 300,000 edit operations (see Table 1). This means that 0.79 % of the text was modified in the correction process (micro-average). The mean modification rate per page (macro-average) is more than ten times higher, which means that a significant portion of the modifications originates from pages with a small amount of text. Both average figures show an exceptionally high standard deviation of 430 %.[25]

A major source for this high variability are errors in Layout Recognition, an early stage of OCR responsible for detecting and ordering blocks of text and other page elements, such as figures and tables. When addressing these errors, the correctors often had to rearrange contiguous spans of text (up to multiple paragraphs) in order to establish the correct reading order. However, corrections that involve moving text regions are not appropriately reflected by tools that focus on local changes on the word or character level: While the actual edit action requires only a few clicks and keystrokes independent of the size of the moved text, the tool records a sequence of

[25] Deleting major portions of a page may lead to an edit distance greater than the length of the corrected text, which results in a modification rate of more than 100 %.

Table 1: Effects of filtering on corpus size and modification rates.

filtering	pages	para-graphs	tokens	characters	cumul. LD	modification rate macro (SD)		micro (SD)	
						%	%	%	%
none	21,246	137,395	6,451,906	37,158,538	293,366	8.51	(430.0)	0.79	(430.1)
page-wise	19,029	110,726	5,857,982	33,886,068	180,987	2.83	(18.9)	0.53	(19.1)
./. DE	17,190	93,575	5,236,748	30,720,939	160,625	2.55	(18.0)	0.52	(18.1)
./. FR	1839	17,151	621,234	3,165,129	22,969	4.36	(23.3)	0.72	(23.6)
para-wise	19,024	107,043	5,838,302	33,773,261	141,592	1.28	(4.5)	0.42	(4.6)
./. DE	17,186	90,855	5,221,461	30,632,199	127,316	1.26	(4.3)	0.42	(4.4)
./. FR	1838	16,188	616,841	3,141,062	14,276	1.39	(5.9)	0.45	(5.9)

deleted characters in one spot and a corresponding insertion elsewhere, measuring a value proportional to the number of characters shifted.

In order to avoid this distorting effect and to measure the amount of typical corrections in running text more reliably, we removed pages that were prone to artificially enlarge the number of edit operations for the evaluation. In particular, we removed table-of-content pages (which had been manually corrected in the initial digitization phase already) and pages with large tables or page-size images. Furthermore, we discarded pages written in one of the sparsely represented languages, i. e. languages other than French or German, and pages containing more than one language,[26] which enabled us to analyze the modifications separately for the two major languages.

Identifiers embedded in the HTML markup allowed us to easily align both versions at the paragraph level. We removed paragraphs that were missing in one of the snapshots (which means that they were either completely deleted or inserted in the correction phase). After this filtering, we were left with a set of around 19,000 pages with a total of 111,000 paragraphs and 33.9 million characters, resulting in a reduction of approximately 10 % (see the rows concerning page-wise filtering in Table 1).

In this filtered corpus, the modifications in the French and German sentences affect 180,000 characters, which equals to an overall modification rate of 0.54 % (micro-average). For French the modification rate is 0.72 %, which is considerably higher than for German (0.52 %). The mean modification rate of all paragraphs (macro-average) shows an even more substantial difference between German and French.

The difference between micro- and macro-average as well as the large variance indicate that, still, a small number of paragraphs have a high modification rate. Inspecting such cases revealed that many occurrences of text reorganisation had remained, e. g. in tables that had not been removed in the first filtering method. Since our ID-based paragraph alignment does not capture text regions moved across paragraphs, we decided to exclude these cases in an additional filtering step. We removed all paragraphs which showed a

[26] As identified by our downstream processing pipeline.

Table 2: Word error rate of the paragraph-wise filtered corpus.

	macro (SD)		micro (SD)	
	4.23 %	(13.25 %)	1.58 %	(13.51 %)
./. DE	4.39 %	(13.50 %)	1.59 %	(13.78 %)
./. FR	3.09 %	(11.59 %)	1.54 %	(11.69 %)

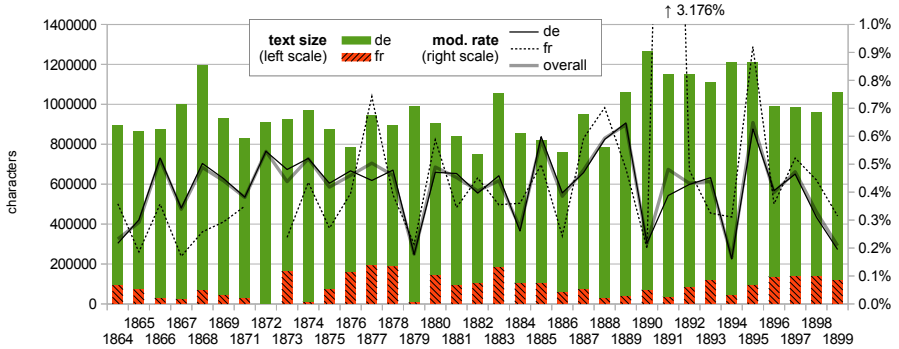

Figure 3: Text size and modification rate (micro-average) in the paragraph-wise filtered data.

change in length of 10 % or more between the two snapshots. This reduced the corpus size by only around 0.3 % (see the rows concerning paragraph-wise filtering in Table 1). The overall modification rate decreased by a fifth to 0.42 %, and the gap between German and French became smaller (0.42/0.46 % respectively in micro-average). As a side effect, this filtering also removed spurious paragraphs caused by spots or dirt.

The modification rate is computed on the level of characters, rather than words. This has the advantage that it can be easily derived from the edit distance and that it reflects directly the amount of edit operations (keystrokes) which the correctors performed. Also, it avoids the complexity of different tokenization rules for different languages, and the rate includes modifications that affect non-word characters (punctuation). However, the amount of words that were changed in the correction phase is a meaningful figure too. Therefore, Table 2 shows the proportion of word tokens in the paragraph-wise filtered corpus that were modified by the correctors. We used the *ocrevalUAtion* tool (Carrasco, 2014) to compute the *word error rate* of the original text, as compared to the crowd-corrected version. It is interesting to see that the word error rate is lower for French than for German (clearly for the macro-average), whereas the inverse distribution is found for the character-level modification rate. This suggests that more misrecognized characters are concentrated in fewer words in the French texts.

Figure 3 shows the modification rates across all yearbooks, plotted against the text size. We found no correlation between the size of a yearbook and its modification

rate, nor did we observe a clear tendency over time (correlation age–modification rate). Often, the modification rates for French and German develop in parallel, which seems intuitive given that paper and printing quality as well as the condition of preservation is the same within one multilingual yearbook. French tends to have a stronger amplitude, showing low values for volumes with a low total rate and even much higher values for highly modified volumes. At least partially, the increased variability might be due to the relatively small amount of French texts, which gives more weight to individual outliers. Some of the slumps in the modification rate (e. g. 1890, 1899) can be attributed to correction efforts early in the digitization process, which were carried out using the user interface of the OCR software. This means that, occasionally, the text quality was already considerably improved before exporting into the online correction system, leaving less work to do for the crowd correctors.

A selection of frequent corrections is given in Table 3. All examples are misrecognized tokens that were corrected multiple times in different places. In many cases, the affected word posed increased challenges to the OCR system, in that it is not expected to be found in a dictionary that covers the general vocabulary of contemporary German or French. Often this is due to orthographic and linguistic variation, such as regional (examples 10–14) and historical spelling (13–16, 34–37) as well as outdated morphology (17–19), or because the word belongs to an open class, such as toponyms (20–23, 44–52), and person names (24–27, 38–40). Many errors are related to spurious or missing diacritic marks (11–14, 17–19, 32–37, 43–44, 48–50). Also, non-alphanumeric characters (55–57), superscripts (51–54), and certain letters (e. g. upper-case *R*, see 7–9, 47) are generally badly recognized. Occasionally, French words appear as if the background dictionary of another language was in place during recognition (41–45). Some place names show a spelling alternation adapted to French orthography, whereas they had been recognized in the spelling of their original language (48–50).

From a natural language processing point of view, it is worthwhile to look at cases that are particularly hard to tackle in automated post-correction. As such, many corrections deal with *real-word* errors (3–6, 10–14, 16–21, 38), i. e. tokens that match an existing word, which means that their erroneous nature can only be revealed through their context or by comparison with the facsimile. A similarly tricky issue is dehyphenation, which cannot be performed mechanically in a linguistically unaware fashion (see example 28 vs. 30). We tried to estimate the amount of different real word error types for words (excluding their non-alphanumeric characters such as quotation marks or hyphenation characters). About 19% of all word types that were corrected at least once can be found in the corrected version of our corpus, and therefore, might be considered as real word errors.

Table 4 shows the most frequent edit operations. Most of the top modifications are concerned with fixing word boundaries through insertion and deletion of spaces and hyphens. Some operations had been carried out using the search and replacement interface, such as the global replacement of quotation marks or the removal of squares and bullets. 35.9 % of the modifications are deletions of one or more characters (mostly punctuation and whitespace characters). Many corrections are related to letters with

Table 3: Frequent word corrections.

German

OCR	corr.	
nnd	und	(1)
zn	zu	(2)
sieh	sich	(3)
Ton	von	(4)
Über	über	(5)
lieber	Ueber	(6)
Eichtung	Richtung	(7)
Kichtung	Richtung	(8)
Bedaktion	Redaktion	(9)
Hessen	liessen	(10)
massig	mässig	(11)
Händen	Handen	(12)
Centralcomite	Centralcomité	(13)
Bureau	Büreau	(14)
Thaies	Thales	(15)
grossenteils	grossentheils	(16)
altern	ältern	(17)
Schütze	Schutze	(18)
Schlüsse	Schlusse	(19)
Eimer	Elmer	(20)
Gesehenen	Geschenen	(21)
Unterwaiden	Unterwalden	(22)
Bergeil	Bergell	(23)
Bubi	Dübi	(24)
Imfeid	Imfeld	(25)
111.	Ill.	(26)
Franche	Francke	(27)
Ueber-gang	Uebergang	(28)
all-mälig	allmälig	(29)
Schnee-und	Schnee- und	(30)

French

OCR	corr.	
11	Il	(31)
ä	à	(32)
oü	où	(33)
complètement	complétement	(34)
mètres	mêtres	(35)
privilège	privilége	(36)
sécher	sècher	(37)
Aimer	Almer	(38)
Ford	Forel	(39)
Nsegeli	Nægeli	(40)
Tun	l'un	(41)
Us	Ils	(42)
ä l'etude	à l'étude	(43)
See	Scé	(44)
Mordes	Morcles	(45)
VOfenpass	l'Ofenpass	(46)
Ehône	Rhône	(47)
Lütschine	Lutschine	(48)
Saas-Fee	Saas-Fée	(49)
Palù	Palu	(50)
S*-Bernard	S^{t}-Bernard	(51)
S'-Gothard	S^{t}-Gothard	(52)

language-independent

OCR	corr.	
')	^{1})	(53)
m 8	m^{3}	(54)
-f-	+	(55)
°/o	%	(56)
Va	½	(57)

diacritic marks, which appear to be particularly challenging for OCR in a multilingual corpus. 7.6 % of the observed edit operations differ only by diacritics (e. g. a→ä or vice versa).

3.2 Quality of Corrections

In order to assess the quality of the corrected pages, we decided to carefully validate them using a representative sample. For having a wide base, sampling units should be as small as possible, e. g. words or even single characters. However, proofreading individual characters is tedious, and judging words also often requires additional context. In the trade-off between coverage and sufficient context, we set the unit size to a span of 1–2 printed lines, on which the proofreaders consented that it is considerably less tiring than

Table 4: Most frequent edit operations.

German						French					
freq.	OCR	corr.	freq.	OCR	corr.	freq.	OCR	corr.	freq.	OCR	corr.
13,970	‹ ›	‹›	528	‹a›	‹ä›	2024	‹ ›	‹›	49	‹»›	‹s›
10,006	‹-›	‹›	496	‹Y›	‹V›	701	‹e›	‹é›	45	‹B›	‹R›
7175	‹"›	‹“›	486	‹é›	‹e›	526	‹-›	‹›	44	‹œ›	‹æ›
3669	‹›	‹ ›	461	‹ • ›	‹›	417	‹›	‹ ›	42	‹›	‹.›
2644	‹i›	‹l›	411	‹u›	‹n›	372	‹"›	‹“›	41	‹*›	‹1›
1942	‹e›	‹c›	407	‹ii›	‹ü›	183	‹.›	‹›	39	‹O›	‹0›
1354	‹.›	‹›	383	‹o›	‹ö›	141	‹-›	‹ ›	38	‹V›	‹l'›
1147	‹K›	‹R›	380	‹ii›	‹n›	128	‹ä›	‹à›	38	‹*›	‹t›
1146	‹E›	‹R›	358	‹ti›	‹ü›	126	‹,›	‹›	37	‹- ›	‹›
1079	‹u›	‹ü›	353	‹™›	‹ m›	124	‹e›	‹è›	36	‹,›	‹.›
1070	‹"›	‹›	337	‹0›	‹O›	114	‹n›	‹„›	35	‹. ›	‹›
912	‹,›	‹›	332	‹,›	‹.›	113	‹i›	‹l›	35	‹I›	‹l›
847	‹B›	‹R›	294	‹›	‹.›	112	‹e›	‹c›	34	‹se›	‹æ›
824	‹li›	‹h›	275	‹*›	‹1›	98	‹é›	‹e›	34	‹-›	‹—›
783	‹-›	‹ ›	269	‹a›	‹u›	98	‹■›	‹›	33	‹'›	‹›
782	‹U›	‹ü›	268	‹—›	‹-›	95	‹ • ›	‹›	33	‹è›	‹é›
766	‹n›	‹u›	236	‹«›	‹e›	94	‹'›	‹›	32	‹li›	‹h›
741	‹ö›	‹o›	229	‹ii›	‹u›	91	‹E›	‹R›	32	‹0›	‹O›
722	‹ü›	‹u›	226	‹- ›	‹›	85	‹—›	‹-›	32	‹Y›	‹V›
713	‹'›	‹›	225	‹B›	‹D›	82	‹ii›	‹ü›	32	‹'›	‹t›
655	‹■›	‹›	221	‹tt›	‹ü›	82	‹K›	‹R›	32	‹e›	‹ê›
625	‹ä›	‹a›	218	‹*›	‹›	80	‹'›	‹1›	31	‹.›	‹ ›
604	‹'›	‹1›	214	‹a›	‹n›	76	‹™›	‹ m›	31	‹a›	‹s›
573	‹e›	‹é›	213	‹i›	‹›	73	‹n›	‹u›	27	‹k ›	‹›
533	‹m›	‹rn›	202	‹»›	‹s›	50	‹"›	‹›	27	‹'›	‹ ›

reading randomly sampled words. Therefore, we divided the filtered corpus into snippets with a soft target size of roughly 100 visible characters (117 including whitespace, on average), which was adjusted accordingly to meet word and page boundaries.

For determining the minimal required size of the sample, we regarded the problem as an application of empirical probability of one character being incorrectly recognized. Preliminary investigation suggested that the rate of remaining errors did not exceed 0.1 % on the level of characters. Since the distribution of the two classes (correctly vs. incorrectly recognized) are very skewed (999:1), we chose a narrow error band of ±0.02 %. With a significance level of $p < 0.01$ (and thus $z^2 \approx 6.635$), the minimal required sample size in terms of characters is:[27]

$$\frac{z^2}{0.0002^2} \times 0.001 \times (1 - 0.001) = 165706.54$$

[27] Assuming that recognition errors are independent of each other, so that sampling sequences of contiguous characters yields the same distribution as sampling individual characters.

As this number is close to $\frac{1}{200}$ of the corpus, we divided the corpus into 200 stratified folds and picked one for proofreading. Each fold contained approximately 1440 snippets that were randomly sampled, but with a distribution representative for the entire corpus with regard to yearbook and language.

The selected sample had a size of approximately 25,000 word tokens. It was independently proofread by two German native speakers with good knowledge of French. They were asked to correct the snippets according to the guidelines of the crowd correctors. For each snippet, an appropriately cropped facsimile image was provided for collation.

5 % of the snippets were modified by at least one proofreader. Well over half of the modifications were done by both correctors in agreement, the rest was contributed by either of them in similar parts. When both proofreaders modified the same word, their modifications were always identical, i. e. they never disagreed on how to correct an error, but only on its mere presence. It is most likely that the disagreements arose from varying attentiveness, rather than differences in judgment: Some of the errors just slipped through, as they already had for our crowd correctors.

We sought for a way to quantify the agreement between the correctors with a standard measure. If we assume that the correction of each detected error is unambiguous, the modifications can be modeled as a binary classification task (namely error *detection*) on the word level. Thus, each word in the sample is a data point, comparing the correctors' decisions to change this word or leave it unchanged. Measuring the agreement with Fleiss' κ (Fleiss, 1971) yielded a value of 0.67. Discussions between the proofreaders revealed that the guidelines were not detailed enough concerning whitespace (e. g. spaces separating the integer and fractional part in decimal numbers). By applying appropriate adjudication to these cases, κ raised to 0.73.

We then merged the proofread snippets into a single gold standard. Judging from the κ score, a few errors may have remained undetected, but we expect them not to be more than a handful, as the number of errors found by only one of the proofreaders and missed by the other was small.

By comparing to the gold standard, we measured the spelling quality of the sample as corrected by the online collaborators. In total, the proofreaders corrected 113 characters in 72 words, which means that the crowd-corrected texts achieved a high accuracy of 99.71 % on the level of words and 99.93 % on the level of characters. Qualitatively, most of the remaining errors were hard-to-spot details, such as missing commas or diacritics (e. g. *avance*/*avancé*) or substitutions of similarly looking letters (e. g. *Clnbhütte*/*Clubhütte*, *Generalyersammlung*/*Generalversammlung*).

Based on the accuracy figures, we estimate the proportion of errors removed by the crowd correctors. The modification rate of the filtered corpus tells us that 141,592 characters were edited (see Table 1). Under the assumption that every modification by the correctors effectively contributed to an error correction, this can be considered as the number of removed character errors. Extrapolating the observed rate of remaining errors in the sample (0.07 %[28]) to the entire corpus (33.8 million characters), we can

[28] The exact observed character error rate is $\frac{113}{169619}$.

estimate the amount of remaining errors to be approximately 22,500 character errors. This means that our online collaborators removed a proportion of $\frac{141592}{141592+22500}$ of all errors, which is a reduction rate of 86 %. For the word level, an analogous computation yields an estimated reduction rate of 85 %.

3.3 OCR19thSAC: an OCR Resource for Training and Testing

While correcting the OCR errors in our heritage corpus of German and French texts from the Alpine domain, we created a large OCR ground truth that can be either used as OCR training and testing material, or for optimizing automatic OCR post-correction, or as a resource for lexicon extraction. The estimated word-level accuracy of 99.7 % provides a good basis for evaluating systems that either process the scanned images of the pages or try to improve upon the output of a standard OCR system. We provide the scanned images, the initial snapshot of the extracted text, and the crowd-corrected ground truth (final snapshot).

We distribute the textual portion of our OCR19thSAC corpus in three different versions:

1. Complete multilingual corpus without filtering, page-wise aligned with the scan images. Provided in two variants: text only and text plus image coordinates of word boundaries. The latter is suited for extracting training material for an OCR engine which often requires image snippets of lines and their corresponding ground truth text.

2. Corpus with page-wise filtering (as described in Section 3.1), paragraph-wise aligned across snapshots; suitable for training a post-correction system.

3. Same as 2, but with additional paragraph-wise filtering, that is, without paragraphs that changed more than 10 % (measured in characters) between the two snapshots, also suitable for post-correction training.

All versions are provided as UTF-8 encoded plain text for both snapshots, that is OCR output quality and crowd-corrected quality, under a Creative Commons Attribution 4.0 International License.[29]

Although many project are involved in the digitization of heritage text material and OCR correction, surprisingly few OCR datasets are available in a suitable form for training and testing. In the course of the *IMPACT* project (Tumulla, 2008), a collection of ground-truth texts was created from digitized historical printed texts. The resource is advertised on the *Impact Centre of Competence*'s website,[30] however, it is only accessible to members.

The open-source OCR framework *OCRopus* (Breuel, 2008) could be trained with our resource. More recently, Breuel, Ul-Hasan, Al-Azawi, and Shafait (2013) have shown that an OCR system based on LSTM neural networks is able to outperform commercial

[29] See `http://pub.cl.uzh.ch/purl/OCR19thSAC` for download
[30] `http://www.digitisation.eu/tools-resources/image-and-ground-truth-resources/`

systems even with a rather small training set. Berkeley's GPU-enabled state-of-the-art historical OCR system *Ocular* (Berg-Kirkpatrick & Klein, 2014)[31] can easily be trained for documents with simple one-column book-like layouts and performs substantially better on historical data than commercial off-the-shelf systems.

4 Conclusion

We have shown that interested volunteers can effectively solve annoying OCR quality problems for the scientific community. In our case we were able to recruit volunteers from an inherently interested community that additionally has a long tradition of citizen science. Other success factors that we consider relevant for projects like ours are: (a) simple and concise guidelines, (b) an easy to use user interface with intuitive user interactions, (c) visual aids for quickly moving between the textual representation and its location in the facsimile image, (d) support for different ways of accessing the text and detecting possible errors, for instance, by reading sequentially or by investigating search results, (e) constant feedback about the correction progress on the level of validated pages, (f) personal correction statistics and high score rankings. The latter is needed in order to keep motivation up for the top volunteers who typically show an incredible amount of dedication to the task. The achieved accuracy of 99.93 % on character-level comes close to the performance of double-keying transcription methods.

Acknowledgements

We would like to thank our volunteer correctors who invested a lot of time and engagement into our crowd-correction project. We also thank Adrian Althaus and Matthias Fluor for implementing the Kokos web platform. SC has been supported by the Swiss National Science Foundation under grant CRSII5_173719.

References

Berg-Kirkpatrick, T., & Klein, D. (2014). Improved typesetting models for historical OCR. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 118–123). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/P14-2020` doi: 10.3115/v1/P14-2020

Breuel, T. M. (2008). The OCRopus open source OCR system. In *Proc. spie* (Vol. 6815, p. 68150F-68150F-15). Retrieved from `http://dx.doi.org/10.1117/12.783598` doi: 10.1117/12.783598

Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A., & Shafait, F. (2013, Aug). High-performance OCR for printed English and Fraktur using LSTM networks. In *2013*

[31] `http://nlp.cs.berkeley.edu/projects/ocular.shtml`

12th international conference on document analysis and recognition (p. 683-687). doi: 10.1109/ICDAR.2013.140

Carrasco, R. C. (2014). An open-source OCR evaluation tool. In *Proceedings of the first international conference on digital access to textual cultural heritage (DATeCH '14)* (pp. 179–184). New York, NY, USA: ACM. doi: 10.1145/2595188.2595221

Chrons, O., & Sundell, S. (2011). Digitalkoot: Making Old Archives Accessible Using Crowdsourcing. In *Proceedings of the 2011 AAAI Workshop on Human Computation* (pp. 20–26). Association for the Advancement of Artificial Intelligence (AAAI). Retrieved from `http://cdn3.microtask.com/assets/download/chrons-sundell.pdf`

Deutsches Textarchiv – Qualitätssicherung. (2016). Retrieved from `http://www.deutschestextarchiv.de/dtaq/`

Dunning, T. (1994). *Statistical identification of language* (Tech. Rep. No. CRL MCCS-94-273). Computing Research Lab, New Mexico State University.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. doi: 10.1037/h0031619

Geyken, A., & Gloning, T. (2015). A living text archive of 15th-19th-century German. corpus strategies, technology, organization. In J. Gippert & R. Gehrke (Eds.), *Historical corpora. challenges and perspectives* (p. 165-179). Tübingen: Narr.

Göhring, A., & Volk, M. (2011). The Text+Berg Corpus: An Alpine French-German Parallel Resource. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN 2011).* Montpellier. Retrieved from `http://www.atala.org/taln_archives/TALN/TALN-2011/taln-2011-court-017`

Haaf, S., Wiegand, F., & Geyken, A. (2013, March). Measuring the correctness of double-keying: Error classification and quality control in a large corpus of TEI-annotated historical text. *Journal of the Text Encoding Initiative [Online]*, *4*. Retrieved from `http://jtei.revues.org/739` doi: 10.4000/jtei.739

Haaf, S., Wiegand, F., & Geyken, A. (2014, March). The dta "base format": A tei subset for the compilation of a large reference corpus of printed text from multiple sources. *Journal of the Text Encoding Initiative [Online]*, *8*. Retrieved from `http://jtei.revues.org/1114` doi: 10.4000/jtei.1114

Holley, R. (2009a). How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine*, *15*(3/4). Retrieved from `http://www.dlib.org/dlib/march09/holley/03holley.html`

Holley, R. (2009b). *Many hands make light work: Public collaborative OCR text correction in Australian historic newspapers.* National Library of Australia.

Holley, R. (2010). Crowdsourcing: How and why should libraries do it? *D-Lib Magazine*, *16*(3/4). Retrieved from `http://dx.doi.org/10.1045/march2010-holley` doi: 10.1045/march2010-holley

Holley, R. (2013). *Crowdsourcing text correction and transcription of digitised historic newspapers: a list of sites.* Retrieved 2016/01/05, from `http://rose-holley.blogspot.ch/2013/04/crowdsourcing-text-correction-and.html`

Kettunen, K. (2016). Keep, change or delete? setting up a low resource OCR post-correction framework for a digitized old Finnish newspaper collection. In D. Calvanese, D. De Nart, & C. Tasso (Eds.), *Digital libraries on the move: 11th italian research conference on digital libraries, ircdl 2015, bolzano, italy, january 29-30, 2015, revised selected papers* (pp. 95–103). Cham: Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-41938-1_11 doi: 10.1007/978-3-319-41938-1_11

Koskeniemmi, K., & Haapalainen, M. (1996). GERTWOL – Lingsoft Oy. In R. Hausser (Ed.), *Linguistische Verifikation: dokumentation zur ersten Morpholympics 1994* (pp. 121–140). Tübingen: Niemeyer.

Levenshtein, V. I. (1966, February). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, *10*(8).

Long, F. A. (1993). Electronic composition and the typesetter. *American Journal of Economics and Sociology*, *52*(2), 223–226. Retrieved from http://dx.doi.org/10.1111/j.1536-7150.1993.tb02536.x doi: 10.1111/j.1536-7150.1993.tb02536.x

Piotrowski, M. (2012). *Natural language processing for historical texts* (Vol. 5) (No. 2). Morgan & Claypool. Retrieved from http://dx.doi.org/10.2200/S00436ED1V01Y201207HLT017 doi: 10.2200/S00436ED1V01Y201207HLT017

Reffle, U. (2011). Efficiently generating correction suggestions for garbled tokens of historical language. *Natural Language Engineering*, *17*(2), 265–282. Retrieved from http://www.journals.cambridge.org/abstract_S1351324911000039

Rice, S. V. (1996). *Measuring the Accuracy of Page-Reading Systems* (Doctoral dissertation, University of Nevada, Las Vegas). Retrieved from http://www.cs.olemiss.edu/~rice/rice-dissertation.pdf

Scientific library services and information systems (LIS): DFG practical guidelines on digitisation. (2009). electronic. Retrieved from http://www.dfg.de/download/pdf/foerderung/programme/lis/praxisregeln_digitalisierung_en.pdf

Seidman, M. J., Flanagan, M., Rose-Sandler, T., & Lichtenberg, M. (2016, jul). Are games a viable solution to crowdsourcing improvements to faulty OCR? – the purposeful gaming and BHL experience. *code4lib*, *33*. Retrieved from http://journal.code4lib.org/articles/11781

Simeoni, M. M. J.-A., Mazzei, A., & Kaplan, F. (2014). *Semi-automatic transcription tool for ancient manuscripts.* IC Research Day 2014: Challenges in Big Data, SwissTech Convention Center, Lausanne, Switzerland, June 12, 2014. Retrieved from https://infoscience.epfl.ch/record/199578/files/Semi-Automatic%20Transcription%20Tool%20for%20Ancient%20Manuscripts%20_%20The%20Venice%20Atlas.pdf

Tumulla, M. (2008, July). IMPACT: Improving Access to Text. *Dialog mit Bibliotheken*, *20*(2), 39–41. ((German article))

Vobl, T., Gotscharek, A., Reffle, U., Ringlstetter, C., & Schulz, K. U. (2014). PoCoTo - an open source system for efficient interactive postcorrection of OCRed historical texts. In *Proceedings of the first international conference on digital access to textual cultural heritage DATeCH '14* (pp. 57–61). ACM Press. Retrieved from

http://dl.acm.org/citation.cfm?doid=2595188.2595197

Volk, M., & Clematide, S. (2014, oct). Detecting code-switching in a multilingual Alpine heritage corpus. In M. Diab, J. Hirschberg, P. Fung, & T. Solorio (Eds.), *Proceedings of the first workshop on computational approaches to code switching* (p. 24-33). Doha, Qatar: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W14-39

Volk, M., Marek, T., & Sennrich, R. (2010, August). Reducing OCR Errors by Combining Two OCR Systems. In *ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 61–65). Retrieved from http://dx.doi.org/10.5167/uzh-35259

von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). re-CAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, *321*(5895), 1465–1468. Retrieved from http://science.sciencemag.org/content/321/5895/1465 doi: 10.1126/science.1160379

Wang, S., Wang, M., & Chen, K. (2013). Boosting OCR accuracy using crowdsourcing. In *Human computation and crowdsourcing: Works in progress and demonstration abstracts, an adjunct to the proceedings of the first AAAI conference on human computation and crowdsourcing, november 7-9, 2013, palm springs, ca, USA* (Vol. WS-13-18). AAAI. Retrieved from http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7538